DOCUMENT RESUME

ED 277 716 TM 860 721

AUTHOR Webster, William J.; Olson, George H.
TITLE An Empirical Approach to Identifying Effective Schools.
PUB DATE Apr 84
NOTE 37p.; Paper presented at the Annual Meeting of the American Educational Research Association (65th, New Orleans, LA, April 23-27, 1984).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Academic Achievement; *Achievement Gains; Elementary Secondary Education; Enrollment Trends; *Evaluation Criteria; *Evaluation Methods; Grade Equivalent Scores; Low Achievement; Predictor Variables; Regression (Statistics); *Research Methodology; *School Effectiveness; Student Characteristics; Teacher Characteristics
IDENTIFIERS California Test of Basic Skills Espanol; Dallas Independent School District TX; Effective Schools Research; Iowa Tests of Basic Skills; Tests of Achievement and Proficiency

ABSTRACT

One approach to identifying effective schools defines effectiveness in terms of student achievement in reading, mathematics, and language usage. Exceptional school achievement is indicated by performance above or below the level expected if students were merely to maintain their previous rate of growth. Regression analysis is used to compute prediction equations by grade level and skill area, independent of school identification. Resulting equations are used to compute scores for comparing gains over expectation for students within schools. This approach has several advantages. It controls for systematic influences governing student composition, provides all schools with an opportunity to demonstrate success, is sensitive to differences in resource management, and concurs with practitioners' views of effective schools. Grade equivalent scores were studied for Dallas (Texas) students in grades 2 to 11, on the Iowa Tests of Basic Skills, California Tests of Basic Skills-Espanol, or Tests of Achievement and Proficiency. Four statistical methods for ranking schools were developed and correlated with school demographic variables, teacher characterisitcs, and previous achievement. These procedures, along with student and teacher attendance data, are used to award stipends to teachers in the highest-ranking Dallas schools. (GDC)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED277716

An Empirical Approach to Identifying Effective Schools

William J. Webster and George H. Olson
Dallas Independent School District

Paper presented at the annual meeting of the American Educational Research Association, New Orleans, Louisiana, April 23-27, 1984.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

W. Webster

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☐ This document has been reproduced as received from the person or organization originating it.

X Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

TM 860 121

BEST COPY AVAILABLE

An Empirical Approach to Identifying Effective Schools[1]

William J. Webster and George H. Olson
Dallas Independent School District

In recent years considerable study has been focused upon defining the characteristics of effective schools (Bloom, 1976; Brookover and Lezotte, 1979; Anderson, et. al., 1982; Duffy, 1980; Edmonds, 1979; Dunn, 1981; Weber, 1971; Levine and Stark, 1981; Glenn, 1981; Doyle, 1981; Emer and Anderson, 1980; Stallings, 1980; Rosenshine, 1979; Rosenshine and Furst, 1973; Slavin, 1980; Gage, 1978; Lysakowski and Walberg, 1982; Block, 1970; Peterson, 1979). These works generally have had three factors in common:

(1) they have been limited primarily to efforts involving elementary schools,

(2) they have been limited primarily to low socioeconomic status groups, and

(3) they have used as their criterion variable either status achievement scores or cross-sectional achievement scores interpreted in a longitudinal manner (studies following individual students over periods of time are rare).

Apparently, little progress has been made toward developing empirical models for identifying effective schools. It is essential that substantive investigations toward identifying effective schools be accelerated, particularly in light of recent trends toward merit pay and teacher accountability.

---

[1] Paper presented at the annual meeting of the American Educational Research Association, New Orleans, Louisiana, April 23-27, 1984.

One approach to identifying effective schools is to define and measure effectiveness in terms of student achievement in the basic skill areas of reading, mathematics, and language usage. While other desirable goals and outcomes of public education could be enumerated easily, under today's heightened public awareness these have become only secondary in importance to basic achievement. This, at least, is the assumption by which the study reported in this paper was motivated.

School effects. In this study a school's effectiveness is associated with exceptional student achievement, defined as measured test performance that is above or below that which would be expected if a school did no more or no less than simply maintain the students' previous rate of achievement growth. In general it is reasonable to expect students to continue to achieve at a given rate. Thus when a school's population of students departs markedly from its own pre-established trend or from the more general trend of similarly achieving students throughout the district, this departure is attributed to a school effect. The problem of measuring a school's effect, then, becomes one of establishing the school's students' rates of achievement, setting expected levels of performance based upon these rates, and determining the extent to which its students, on the average, exceeded or fell short of expectation. Essentially, this is the same problem as establishing a school's average level of achievement after controlling for its students' previous levels of achievement. The procedures involve using regression analysis to compute prediction equations by grade level and skill area independent of school identification and then using these equations within schools to obtain mean gains over expectation.

Advantages of the approach. There are a number of advantages that can be listed for this approach. First, it controls for systematic influences governing the student composition of schools. Since, as many studies have shown, test score performance is highly correlated with student background, demographic, and environmental factors, controlling on previous test score performance indirectly controls for these other variables as well. Second, it gives all schools an even opportunity to demonstrate success. Schools need only focus upon accelerating their own students' rates of achievement; they need not compete with each other in terms of absolute achievement levels. Thus, schools derive no particular advantage by being composed of higher- or lower-ability students. Third, since all schools are allocated resources on the basis of specific formulae (schools having similar needs are provided similar resources) the procedure is sensitive to differences in the way resources are managed. Finally, the approach is in consonance with many practitioners' views of what constitutes an effective school.

Limitations. One limitation of this approach is that it focuses on standardized test performance. Certainly, learning and achievement occurs in areas not measured by these tests, and in areas where the tests are less sensitive. Nevertheless, it must be realized that much of the recent public attention directed at the status of achievement in the nation's schools is really directed at the results of standardized testing programs. No matter what schools do to affect achievement the effect of their efforts will be recognized only to the extent that they affect standardized test scores.

A second limitation of this approach is that it does not address causal connections with those factors that contribute to school effectiveness. While it is possible to identify schools in which exceptional achievement has occurred, it is not possible, within the scope of this study, to identify those factors that caused the exceptional achievement. On the other hand, it would make little sense to embark upon a study of causal factors before having a clear definition of what constitutes school effectiveness. The study reported here provides a good, working, operational definition of effectiveness.

Cases and Variables

Cases. Cases included in this study were all Dallas students tested, at grade levels 2 through 11, in spring 1983 who met the following eligibility criteria:

(1) obtained a valid test score on the reading, mathematics, or language arts (or written expression) subtest of the Iowa Tests of Basic Skills (ITBS), California Tests of Basic Skills-Espanol (CTBS Espanol), or Tests of Achievement and Proficiency (TAP) batteries;

(2) had valid test scores in the same tested areas in both the preceding two years[2,3] (spring 1982 and spring 1981); and

---

[2] Second grade students tested in spring 1983 required only the previous year's data.

[3] In a preliminary step to this study regression equations were computed using up to four years of historical achievement data. The inclusion of each additional year's data resulted in a net reduction of roughly 15 percent of the surviving population. At the same time the increment in variance explained by the addition of each year after the second was less than one percent (the addition of the second year's data generally accounted for an increase in predicted variance of three to seven percent). Thus two years of historical data seemed sufficient both to conserve population sizes and to attain adequate prediction.

(3) did not meet any of the exclusion criteria normally used for excluding cases from aggregation in systemwide achievement reports (thus certain classifications of special education students and other severely skill-deficient students were excluded from consideration).

Over the last several years testing has been administered annually in the spring, the ITBS at grade levels 1 through 8 and the TAP at grade levels 9 through 11. Significant numbers of CTBS Espanol tests have been administered only at grade levels 1 through 3. For students at grade levels 9 and 10, previous years' scores could have been obtained from the ITBS. However, at lower grade levels, if a student was administered the CTBS Espanol in spring 1983 he or she had also to have been tested with the CTBS in all prior years to be included here. Similarly, students administered the ITBS in spring 1983 but the CTBS in previous years were excluded. The results of this selection procedure are summarized in Table 1, where the numbers of cases in each group identified by grade level, subtest, and battery are given.

See Table 1

Data. The data were grade-equivalent (GE) scores in reading (reading total), mathematics (math total), and language arts (language total (ITBS), or written expression (TAP)). Only the reading and mathematics scores from the CTBS-Espanol were used since these were the only subtests administered on a regular basis. On all three batteries the GE score is derived as a normalized growth scale.

Table 1

Number of Cases Entering Regression Equation for Each Grade Level, Subtest, and Battery Classification

Grade	ITBS			CTBS		TAP		
	R	M	L	R	M	R	L	M
2	7574	7367	5188	634	632	—	—	—
3	6501	6263	6300	166	163	—	—	—
4	6932	6768	6691	57	56	—	—	—
5	6703	6526	6604	—	—	—	—	—
6	7132	6954	7039	—	—	—	—	—
7	7107	6915	6928	—	—	—	—	—
8	6737	6602	6598	—	—	—	—	—
9	—	—	—	—	—	6707	6532	6347
10	—	—	—	—	—	5255	5212	5074
11	—	—	—	—	—	4989	4971	4870

Note. R = Reading; M = Mathematics; L = Language Arts (ITBS), Written Expression (TAP).

Regression Phase

The first phase of this study involved computing 36 prediction equations, one for each of the classifications given in Table 1. In each equation the criterion was the spring '83 subtest score and the predictors (one at grade 2; two elsewhere) were the previous years' subtest scores.

Accuracy of prediction was measured by

(1) standard errors of measurement,

(2) coefficients of determination, and

(3) indices of forecasting efficiency.

For each equation these indices are given in Tables 2, 3, and 4.

See Tables 2, 3, and 4

It is obvious, from an inspection of Tables 2, 3, and 4 that prediction generally improved with increasing grade level. This finding by itself is interesting for it suggests that with increasing grade level achievement becomes more and more dependent upon earlier years' achievement. Thus at higher grade levels schools have a lesser opportunity to show a differential effect on student achievement.

Prediction and Estimation Phase

The next phase of this study involved computing expected scores and differences between expected scores and actual scores for students within schools. This was done separately by subtest within grade level and then

Table 2

Standard Errors of Estimate for Each Grade Level, Subtest, and Battery Classification

Grade	ITBS			CTBS		TAP		
	R	M	L	R	M	R	M	L
2	.88	.70	1.03	.79	.66	—	—	—
3	.92	.67	.92	.88	.87	—	—	—
4	.94	.73	.88	1.29	.65	—	—	—
5	.90	.70	.79	—	—	—	—	—
6	.95	.70	.83	—	—	—	—	—
7	.96	.68	.79	—	—	—	—	—
8	1.00	.70	.84	—	—	—	—	—
9	—	—	—	—	—	1.73	1.59	1.91
10	—	—	—	—	—	1.80	1.52	1.92
11	—	—	—	—	—	1.96	1.60	2.00

Note. R = Reading; M = Mathematics; L = Language Arts (ITBS), Written Expression (TAP).

Table 3

Coefficients of Determination (R^2) for Each Grade Level, Subtest, and Battery Classification

Grade	ITBS			CTBS		TAP		
	R	M	L	R	M	R	M	L
2	.35	.28	.33	.21	.25	—	—	—
3	.43	.39	.39	.31	.23	—	—	—
4	.57	.57	.62	.12	.32	—	—	—
5	.72	.72	.76	—	—	—	—	—
6	.73	.77	.80	—	—	—	—	—
7	.77	.83	.86	—	—	—	—	—
8	.79	.84	.87	—	—	—	—	—
9	—	—	—	—	—	.71	.71	.65
10	—	—	—	—	—	.78	.78	.70
11	—	—	—	—	—	.78	.80	.72

Note. R = Reading; M = Mathematics; L = Language Arts (ITBS), Written Expression (TAP).

Table 4

Indices of Forecasting Efficiency for Each Grade Level, Subtest, and Battery Classification

Grade	ITBS			CTBS		TAP		
	R	M	L	R	M	R	M	L
2	19.17	15.10	18.43	11.32	12.31	———	———	———
3	24.55	21.64	21.69	16.51	11.76	———	———	———
4	34.56	34.10	41.40	4.68	15.99	———	———	———
5	46.80	46.97	50.66	———	———	———	———	———
6	48.88	52.46	55.16	———	———	———	———	———
7	52.56	59.23	62.86	———	———	———	———	———
8	54.20	60.16	63.29	———	———	———	———	———
9	———	———	———	———	———	46.21	45.98	40.53
10	———	———	———	———	———	52.98	53.03	45.57
11	———	———	———	———	———	53.12	55.17	46.87

Note. R = Reading; M = Mathematics; L = Language Arts (ITBS), Written Expression (TAP).

aggregated over subtests and grade levels within schools. Specific details are given below.

Individual difference scores. Let

Y_{sgi} be the spring 1983 GE score for individual i at grade level g on subtest s, and

X_{1sgi}, X_{2sgi} be the spring 1981 and spring 1982 subtest scores for the same individual.

Then the expected 1983 score on s for this individual was given by the linear function

$$\hat{Y}_{sgi} = f_t (X_{1sgi}, X_{2sgi})$$

where the function, f_t, was the prediction equation corresponding to the appropriate grade-level, subtest, and battery (t) classification as given, for instance, in Table 2. Individual difference scores were then computed as

$$d_{sgi} = Y_{sgi} - \hat{Y}_{sgi}$$

Individually standardized difference scores. The d_{sgi} eventually were to be aggregated over subtests and grade levels within schools. However, within a group to be aggregated both the reliability and the scale of the individual d_{sgi} varied depending upon the particular function used to compute $\hat{Y}_{sgi}$ and the individual's location in the domain of the predictors. For instance, an expected grade-3 score in reading computed by the function derived from CTBS-Espanol data was less reliable than a similar expectation based upon ITBS data. Similarly, expectations at higher grade levels were generally more reliable than expectations at lower grade levels. Additionally, the expectations for individuals close to the centroid of the predictors were more reliable

than those for individuals further away. These differences in reliability of the $\hat{Y}_{sgi}$ in turn affected the scale of the d_{sgi}. A unit difference in grade-3 reading on the CTBS-Espanol, for instance, was not comparable to a unit difference in grade-3 reading on the ITBS.

To correct for differences in scale and reliability the individual d_{sgi} were standardized by dividing by their individual standard errors of prediction. This yielded individually standardized scores,

$$\delta_{sgi} = d_{sgi} / \sqrt{(1/nS^2 + D_s R_{xx} D_s)}$$

where

S^2 = residual mean square,

D_s = diag matrix of predictor standard deviations,

R_{xx} = predictor correlation matrix, and

n = number of observations in the classification from which the prediction equation was derived.

Aggregation across grade levels and subtests. After d_{sgi} and δ_{sgi} were computed on each subtest for each individual in every school, they were then aggregated, first across subtests and then over grade levels within each school. In the Dallas school district there are five types of schools identified by their grade-level configurations:

primary centers, K-3

intermediate centers, 4-6;

elementary schools, K-6;

middle schools, 7-8; and

high schools, 9-12.

Thus, the number of grade levels aggregated within a school depended upon the type of grade-level configuration.

Statistics used in ranking schools. Several statistics, based on the d's and δ's, were computed from the aggregated data for each school. These are defined as follows.

Mean Difference Score:

$$\bar{d} = \sum_{sgi}\sum\sum d_{sgi}/N,$$

the simple mean across all $N = \sum_{sg}\sum n_{sg}$ subtests and grade levels.

Weighted Mean Standard Difference Score:

$$\bar{w}_d = (\bar{d}_{sg}/\sigma_{\bar{d}_{sg}})/n_{sg}$$

where the mean differences for each subtest by grade-level classification were each divided by their own standard error of the mean and then averaged over grade levels and subtests.

Pooled Mean Standard Difference Score:

$$\bar{z}_d = \bar{d} \,/\, \sigma_{\bar{d}}$$

where the standard error of the mean is based on the pooled observations over grade levels and subtests.

Similar statistics were computed from the δ's.

Mean Individually Standardized Difference Score:

$$\bar{\delta} = \sum_{sgi}\sum\sum \delta_{sgi}/N.$$

Weighted Mean Standard Individually Standardized Difference Score:

$$\bar{w}_\delta = \sum_{sg}\sum(\bar{\delta}_{sg}/\sigma_{\bar{\delta}_{sg}})/n_{sg}.$$

Pooled Mean Standard Individually Standardized Difference Score:

$$\bar{z}_\delta = \bar{\delta} \,/\, \sigma_{\bar{\delta}}.$$

In addition to the six statistics described above, two additional statistics were computed in an effort to take into account the

variability of mean estimates. These were called Lower-Bound d (LB_d) and Lower-Bound δ (LB) and are defined as follows:

$$LB_{\bar{d}} = \bar{d} - \sigma_{\bar{d}}, \text{ and}$$

$$LB_{\bar{\delta}} = \bar{\delta} - \sigma_{\bar{\delta}}.$$

School Ranking Phase

The final phase of the study involved ranking the schools on each of the preceding eight statistics. Ranking was done separately within type of school identified by type of grade-level configuration. There were

38 K-3 schools,

7 4-6 schools,

79 K-6 schools,

23 7-8 schools, and

25 9-12 schools.

Results

Rank distributions. The rank distributions are summarized for each ranking statistic in Table 5. Excluding the lower-bound statistics,

See Table 5

the statistics yielding the narrowest apparent spread among schools were the simple means, $\bar{d}$ and $\bar{\delta}$. The widest spread was provided by $\bar{z}_d$ and $\bar{z}_\delta$.

$\bar{d}$ was a measure of a school's average gain over expectation, while $\bar{\delta}$ was the average individualized gain expressed in standard units of individual conditional expected distributions. Neither statistic takes into account the variance of the gain (d_i or δ_i) within school. Thus, schools having larger variances or smaller populations have less

Table 5

Distribution of Ranking Statistics by Type of School (Grade Configuration)

		Grade Configuration					
		K-3	4-6	K-6	7-8	9-12	All
	Max	.62	.31	1.01	.35	.88	1.01
$\bar{d}$	Mean	- .02	.19	.04	.04	.03	.02
	Min	- .55	- .22	- .66	- .22	- .33	- .66
	Max	6.30	2.53	7.19	4.45	7.16	7.19
$\bar{w}_d$	Mean	- .56	2.08	.12	.42	- .00	- .01
	Min	- 6.38	-2.96	- 4.70	- 6.28	- 9.28	- 9.28
	Max	9.49	6.44	19.23	10.01	18.65	19.23
$\bar{z}_d$	Mean	- .86	5.72	1.10	1.08	.26	.53
	Min	-11.89	-7.63	-12.51	-14.78	- 7.05	-14.78
	Max	1.47	.69	2.35	.87	.99	2.35
$\bar{\delta}$	Mean	- .05	.46	.07	.08	.02	.04
	Min	- 1.27	- .60	- 1.47	- .53	- .39	- 1.47
	Max	6.30	2.95	8.23	4.83	7.48	8.23
$\bar{w}_\delta$	Mean	- .56	2.41	.09	.33	.07	- .05
	Min	- 6.38	-3.47	- 4.71	- 6.38	- 3.02	- 6.38
	Max	9.57	7.05	20.65	10.81	19.24	20.65
$\bar{z}_\delta$	Mean	- .89	6.58	.89	.84	.03	.33
	Min	-11.93	-9.34	-13.44	-14.99	- 7.30	-14.99
	Max	.55	.25	.96	.32	.83	.96
IB_d	Mean	- .07	.18	.00	.02	.01	- .01
	Min	- .61	- .25	- .78	- .24	- .40	- .78
	Max	1.32	.60	2.24	.79	.94	2.24
IB_δ	Mean	- .16	.45	- .01	- .02	- .02	- .04
	Min	- 1.40	- .67	- 1.71	- .57	- .47	- 1.71

stable (reliable) means. For this reason, $\bar{d}$ and $\bar{\delta}$ are of interest merely for descriptive purposes. They indicate, in an absolute sense, the extent to which each school exceeded or fell short of its expectation. They are only roughly comparable across schools.

Table 5 shows that the largest average gain over expectation was 1.01 GE units, attained by one of the K-6 schools. Although the table does not indicate so, this same school also registered the largest average individualized gain (2.35 standard units). The largest average negative gain, -.66 GE units, also occurred in a K-6 school. Again, this school also had the largest average negative individualized gain (- 1.47 standard units).

Given the way in which the statistics were computed, the districtwide gain over expectation should have been zero. Thus, $\bar{d}$ and $\bar{\delta}$, over all schools, were .02 and .04 — very close to their expected values.

The remaining statistics take the variances of the d_i and δ_i and population sizes, within schools into account. the $\bar{w}$'s and $\bar{z}$'s divide the $\bar{d}$'s and $\bar{\delta}$'s by their respective, school-specific standard errors of the mean. The lower-bound statistics bracket one standard error of the mean below $\bar{d}$ and $\bar{\delta}$. Therefore, these statistics are all comparable across schools.

The main difference between the $\bar{w}$'s and $\bar{z}$'s is in the way in which their respective standard errors of the mean were computed. The standard errors used in deriving $\bar{w}_d$ and $\bar{w}_\delta$ were computed from a weighted average of the component grade-level by subtest variances and population sizes, whereas the standard errors for deriving $\bar{z}_d$ and $\bar{z}_\delta$ were computed from the pooled variance and population size over all subtests and grade levels.

Table 5 shows that the largest standardized gain ($\bar{w}_d$ or $\bar{z}_d$) occurred in a K-6 school (7.19 and 19.23, respectively) — the same school registering the largest absolute gain ($\bar{d}$). The largest standardized individualized gain ($\bar{w}_\delta$ or $\bar{z}_\delta$) also occurred in a K-6 school (actually, the same school). Again, this was the school registering the largest individualized gain.

The lower-bound statistics (LB_d and LB_δ) allowed probability statements to be made regarding the placement of the individual schools in the ranked distributions. Given the population sizes of each school, the probability that the true $\bar{d}$ (or $\bar{\delta}$) lies above LB_d (or LB_δ) is always greater than .8. Of course, a more conservative lower bound could have been chosen.

Correlations among ranking statistics. The correlations of the eight ranking statistics with each other are given in Table 6. All

See Table 6

correlations are extremely high, whether computed across schools within type of school or over all schools. This should not be considered a remarkable finding given the way in which the statistics were computed.

While the correlations are all high, differences among correlations, over all schools, of at least .05 units are significant well beyond the .05 level of probability (see Ferguson, 1966, p. 188 for a test of the significance between correlation coefficients in correlated samples). Thus the following generalizations can be made. The simple means, $\bar{d}$ and $\bar{\delta}$, tend to correlate lowest with $\bar{w}$'s and $\bar{z}$'s — an indication that the standardized ranking statistics significantly affected the ordering of

Table 6

Correlations among Ranking Statistics

	Ranking Statistics						
	$\bar{d}$	$\bar{w}_d$	$\bar{z}_d$	$\bar{\delta}$	$\bar{w}_\delta$	$\bar{z}_\delta$	LB_d
K-3 Schools (N = 38)							
$\bar{w}_d$	969[a]						
$\bar{z}_d$	964	988					
$\bar{\delta}$	999	971	964				
$\bar{w}_\delta$	969	999	988	971			
$\bar{z}_\delta$	964	989	999	965	989		
LB_d	997	964	958	996	964	959	
LB_δ	995	966	958	997	965	961	999
4-6 Schools (N = 7)							
$\bar{w}_d$	987						
$\bar{z}_d$	972	996					
$\bar{\delta}$	999	997	973				
$\bar{w}_\delta$	992	998	993	993			
$\bar{z}_\delta$	978	996	999	980	996		
LB_d	999	992	980	999	996	985	
LB_δ	999	991	979	999	996	985	999
K-6 Schools (N = 79)							
$\bar{w}_d$	972						
$\bar{z}_d$	933	971					
$\bar{\delta}$	998	977	941				
$\bar{w}_\delta$	974	998	971	980			
$\bar{z}_\delta$	930	971	998	941	973		
LB_d	998	970	924	996	971	920	
LB_δ	997	976	933	998	977	933	998

(continued on next page)

Table 6 (continued)

Correlations among Ranking Statistics

	Ranking Statistics						
	$\bar{d}$	$\bar{w}_d$	$\bar{z}_d$	$\bar{\delta}$	$\bar{w}_\delta$	$\bar{z}_\delta$	LB_d
K-3 Schools (N = 38)							
$\bar{w}_d$	908[a]						
$\bar{z}_d$	904	998					
$\bar{\delta}$	993	919	912				
$\bar{w}_\delta$	919	998	995	934			
$\bar{z}_\delta$	915	999	997	929	999		
LB_d	995	927	924	997	939	936	
LB_δ	983	928	922	998	944	939	995
9-12 Schools (N = 25)							
$\bar{w}_d$	966						
$\bar{z}_d$	968	997					
$\bar{\delta}$	998	972	973				
$\bar{w}_\delta$	966	999	996	974			
$\bar{z}_\delta$	968	997	999	975	998		
LB_d	994	973	976	997	974	976	
LB_δ	988	974	976	995	976	978	998
All Schools (N = 172)							
$\bar{w}_d$	942						
$\bar{z}_d$	931	967					
$\bar{\delta}$	970	940	922				
$\bar{w}_\delta$	946	998	971	944			
$\bar{z}_\delta$	928	962	997	923	969		
LB_d	996	942	930	970	945	926	
LB_δ	966	937	918	997	941	918	972

[a]Decimal points are suppressed.

schools. On the other hand, the simple means tended to correlate highest with their respective lower bounds. This latter finding can possibly be explained by the fact that schools ranking toward the extremes tended to have relatively homogeneous populations in terms of achievement gains. The $\bar{w}$ and $\bar{z}$ statistics tended to correlate lowest with the lower-bound statistics and about the same with each other. As a result ranking based on the standardized statistics, the $\bar{w}$'s and $\bar{z}$'s, were significantly different from those based on the lower-bound statistics. It should be noted, however, that while the school rankings differed according to which ranking statistic was used, in a practical sense the different rankings were not particularly appreciable. Nevertheless, if the ultimate goal is to rank schools according to effectiveness according to some single criterion, a choice among the statistics has to be made.

Preferred ranking statistics. There appeared to be little empirical basis on which to choose one statistic over the others. The simple means could be rejected because, as previously discussed, their reliabilities varied across schools. The $\bar{w}$ and $\bar{z}$ statistics were equal candidates for consideration. However, for no better reason than the fact that the $\bar{z}$'s yielded greater differentiation among schools, the $\bar{z}$'s were preferred over the $\bar{w}$'s. The lower-bound statistics were considered good candidates for selection because of the information they provided concerning the probability of the location of the means. Thus the choice of ranking statistics was narrowed to four: $\bar{z}_d$, $\bar{z}_\delta$, LB_d, and LB_δ. Additional data were then used to help narrow the choice even further.

Correlation with enrollment variables. Data were available on several school enrollment and attendance variables:

student enrollment,

percent white students,

percent black students,

percent Hispanic students,

percent of students receiving Chapter I services,

percent of students in free lunch programs, and

the student average daily attendance.

Correlations of the four preferred ranking statistics with these enrollment and attendance variables are given in Table 7, by type of school and for all schools. Only a few of these are significant at a .05

See Table 7

level of probability. Among those that were significant, systematic patterns occurred within the seven 4-6 schools, where the percent Hispanic students was negatively correlated with all four preferred ranking statistics, and among the K-6 schools where all the statistics were correlated with percent black and average daily attendance. The high negative correlation with percent Hispanic in the 4-6 schools was apparently due to two schools, both of which had considerably higher percentages of Hispanic students than the remaining schools, and both of which ranked at the bottom of the distribution. Although this situation has not been investigated further, it is possible that many of these students were previously in bilingual programs, but because of a provision in the program were required to be tested in English even though they may not yet have become facile with the level of language development required by the ITBS. Of course, the correlation may simply be spurious.

Table 7

Correlations of Preferred Ranking Statistics and Selected Enrollment and Attendance Statistics

Preferred Ranking Statistic	Enrlmnt	% White	% Black	% Hisp	%Chap I	% Free Lnch	ADA
			K-3 Schools (N = 38)				
$\bar{z}_d$	-176[a]	306	-320	092	-287	-287	027
$\bar{z}_\delta$	-163	297	-311	091	-260	-280	039
LB_d	-069	312	-321	096	-265	-307	073
LB_δ	-067	309	-316	093	-262	-305	088
			4-6 Schools (N = 7)				
$\bar{z}_d$	-378	-063	249	-904*	——	-125	421
$\bar{z}_\delta$	-402	-058	242	-893*	——	-122	396
LB_d	-521	-069	237	-814*	——	-105	337
LB_δ	-518	-057	224	-809*	——	-110	317
			K-6 Schools (N = 79)				
$\bar{z}_d$	005	-205	289*	-169	——	-122	310*
$\bar{z}_\delta$	006	-214	304*	-180	——	-129	323*
LB_d	079	-224	315*	-176	——	-124	251*
LB_δ	082	-229*	332*	-193	——	-129	268*
			7-8 Schools (N = 23)				
$\bar{z}_d$	-243	204	-002	-199	——	-270	275
$\bar{z}_\delta$	-239	171	040	-228	——	-270	312
LB_d	-360	097	056	-184	——	-231	403
LB_δ	-323	072	108	-227	——	-252	422*

(continued on next page)

Table 7 (continued)

Correlations of Preferred Ranking Statistics and Selected Enrollment and Attendance Statistics

Preferred Ranking Statistic	Enrlmnt	% White	% Black	% Hisp	%Chap I	% Free Lnch	ADA
			9-12 Schools (N = 25)				
$\bar{z}_d$	008[a]	185	-180	019	——	-013	-137
$\bar{z}_\delta$	002	166	-160	013	——	000	-142
LB_d	023	106	-102	004	——	-041	-067
LB_δ	035	082	-081	001	——	-019	-087
			All Schools (N = 172)				
$\bar{z}_d$	-023	071	002	-073	——	-162	107
$\bar{z}_\delta$	-029	052	025	-085	——	-154	116
LB_d	016	055	012	-065	——	-174*	089
LB_δ	024	040	042	-089	——	-195*	092

[a]Decimal points are suppressed.

*$p \leq .05$

The significant correlations with percent black in K-6 schools present a potential problem. There are really two distinct classes of schools represented by the K-6 grade configuration: schools that receive students in grades 4-6 under a court-ordered desegregation busing plan and schools that are exempt from this plan either because they are considered already naturally integrated or because they are majority black but are exempt under a special provision in the plan. For the most part schools that receive students under the busing plan have, when all six grade levels are taken into consideration, smaller percentages of black students than do the other K-6 schools. Consequently, this finding could suggest that the two classes of K-6 schools should be formally distinguished and rankings performed separately within each group. The same explanation also might apply to the significant correlation with average daily attendance. In general, students (of all races) being bused under the desegregation plan exhibit poorer attendance records.

Significant differences between school types on the same correlated variables were also observed. Thus the correlations with percent white were positive in K-3 schools but significantly different and negative in K-6 schools. On the other hand, correlations with percent black were negative in K-3 schools but significantly different and positive in both 4-6 and K-6 schools. Again, the answer could lie with the district's desegregation program. The 4-6 level is the first level at which court-ordered desegregation busing occurs. The higher the percentage black, the less likely that the school is a receiving center. Thus this finding tends to support earlier district studies that have shown that black children left in neighborhood schools through grade 6 perform

better at the 4-6 level than do black children who are bused (Sheehan and Marcus, 1978).

Correlations with teacher variables. In addition to enrollment and attendance variables, data were available on several teacher-related variables within each school:

percent male teachers,

percent white teachers,

percent black teachers,

percent hispanic teachers,

median years experience,

teacher rates of absenteeism, and

pupil/teacher ratios.

Correlations of the four preferred ranking statistics and these variables are given in Table 8. Among the many correlations given the incidence

See Table 8

of significant correlation was sparse. The only systematic pattern, if indeed it could be called a pattern, occurred within K-6 schools where both lower-bound statistics correlated negatively with percent male. It generally can be concluded from these correlations that the current distribution of measurable teacher resources had little influence on school rankings.

Correlations with concurrent and previous years' achievement. Correlations between the preferred ranking statistics and school achievement means in 1982 and 1983 are given in Table 9. School means

See Table 9

Table 8

Correlations of Preferred Ranking Statistics and Selected Teacher Variables

	Preferred Ranking Statistic							
	$\bar{z}_d$	$\bar{z}_\delta$	LB_d	LB_δ	$\bar{z}_d$	$\bar{z}_\delta$	LB_d	LB_δ
	K-3 Schools (N = 38)				4-6 Schools (N = 7)			
% Male	-066[a]	-055	-083	-073	-731	-724	-633	-643
% White	076	081	068	072	222	256	257	282
% Black	-095	-094	-092	-084	-056	-080	-088	-112
% Hisp	019	015	023	013	-525	-541	-518	-517
Medexp	-018	-017	010	022	-633	-653	-647	-657
Absent	030	021	-027	-030	674	673	625	631
P/T Ratio	-262	-256	-208	-198	214	-193	-171	-170
	K-6 Schools (N = 79)				7-8 Schools (N = 23)			
% Male	-238	-233	-312*	-308*	223	253	291	326
% White	030	030	011	112	215	178	085	049
% Black	129	140	130	146	-176	-134	-122	-058
% Hisp	-135	-144	-179	-193	-014	-044	017	-042
Medexp	017	013	-015	011	320	319	328	335
Absent	-026	-025	002	-004	014	033	229	226
P/T Ratio	080	072	089	086	-166	-167	-431*	-382
	9-12 Schools (N = 25)				All Schools (N = 172)			
% Male	-096	-095	-200	-182	-034	-041	-046	-020
% White	235	215	132	122	124	116	113	097
% Black	-310	-287	-227	-210	-068	-053	-046	-002
% Hisp	306	289	344	321	-041	-052	-057	093
Medexp	221	222	218	226	067	055	053	047
Absent	-294	-290	-212	-205	001	004	002	011
P/T Ratio	-091	-089	-101	-085	-062	-056	-077	-081

[a]Decimal points are suppressed.

*$p \leq .05$

Table 9

Correlations of Preferred Ranking Statistics with Previous and Current Year's Mean Achievement Levels

Preferred Statistic	Reading '82	Reading '83	Mathematics '82	Mathematics '83	Language '82	Language '83
K-3 Schools (N = 38)						
$\bar{z}_d$	230	514*	333*	255*	155	551*
$\bar{z}_\delta$	224	511*	318	261*	148	561*
LB_d	246	508*	304	281*	173	550*
LB_δ	245	504*	295	278*	174	556*
4-6 Schools (N = 7)						
$\bar{z}_d$	370	447	404	512	399	498
$\bar{z}_\delta$	376	449	408	520	402	502
LB_d	546	605	580	688	568	650
LB_δ	507	566	542	655	528	615
K-6 Schools (N = 79)						
$\bar{z}_d$	136	297*	117	257*	168	324*
$\bar{z}_\delta$	130	287*	110	257*	161	320*
LB_d	131	272*	136	239*	160	314*
LB_δ	127	268*	128	240*	156	315*
7-8 Schools (N = 23)						
$\bar{z}_d$	231	282	183	291	204	306
$\bar{z}_\delta$	194	238	139	253	170	272
LB_d	181	213	116	236	153	261
LB_δ	076	100	009	133	056	162
9-12 Schools (N = 25)						
$\bar{z}_d$	152	323	146	464*	149	394
$\bar{z}_\delta$	141	310	137	457*	138	385
LB_d	182	342	180	488*	188	437*
LB_δ	145	308	147	462*	152	409*

Table 9 (continued)

Correlations of Preferred Ranking Statistics with Previous and Current Year's Mean Achievement Levels

Preferred Statistic	Reading		Mathematics		Language	
	'82	'83	'82	'83	'82	'83
All Schools (N = 172)						
$\bar{z}_d$	131	172*	119	167*	141*	208
$\bar{z}_\delta$	115	157	103	154	124*	193
LB_d	137	188*	124	187*	143*	231
LB_δ	128	156	117	153	132*	191

[a]Decimal points are suppressed.

*$p \leq .05$

for 1982 are for the same populations tested in 1983. Before turning to these correlations, however, it is useful to consider what would be expected if the ranking statistics performed as desired. For instance, it is desirable to have the ranking statistics be independent of students' ability level as measured by prior achievement history. Since the statistics are computed from residual gains, differences among students due to prior history have already been controlled.

Another desirable property is that the ranking statistics be uncorrelated with schools' mean achievement for the previous year. An unfair advantage would exist if schools that began the year with higher-scoring students tended always to rank higher than schools that began with poorer-scoring students. The correlations with 1982 achievement, given in Table 9, demonstrate that the statistics performed as expected. Except for the one instance of a significant correlation between $\bar{z}_d$ and mathematics achievement in K-3 schools, ranking by the statistics was uncorrelated with beginning achievement. This finding was particularly important for it, more than any other finding, clearly establishes the fairness of the procedures.

It is tempting also to desire the ranking statistics to be independent of schools' concurrent levels of achievement. The statistics would have limited utility if those schools with higher performing students always ranked higher than schools with poorer performing students. While the statistics would still provide a standardized measure of effectiveness they would have no affect on the ranked standings. Therefore, high correlations between the ranking statistics and school achievement levels could easily jeopardize the usefulness of

the statistics for ranking purposes. On the other hand, it is not always reasonable to expect these correlations to vanish. If schools had their students for only one year, then the correlations probably should vanish. However, schools generally hold most of their students over several years. It is reasonable then to expect the more effective schools to have a greater, cumulative, positive impact on their students' achievement. The converse would hold for schools that are less than effective. Thus, while very high correlations are not desirable, there is no reason to expect very low correlations, although these would not denigrate the use of the statistics.

In light of the preceding discussion the correlations given in Table 9 conform to expectation. The highest correlations occur in K-3 schools. This can be attributed to two factors: (1) K-3 schools hold most of their students over four years, and (2) achievement in these schools is determined less by previous years' achievement than it is in schools of higher grade level configurations. Here the potential for impact over several years is greater than in other types of schools.

The correlations in K-6 schools are also moderately high (all are significant). In these schools the cumulative effect of long-term impact perhaps is most apparent. What differences that exist between the two classes of K-6 schools (discussed earlier) have not yet been investigated.

The correlations for 4-6 schools are all high but not significant due to the small number of 4-6 schools. That the correlations for 7-8 schools are not significantly different from zero could be interpreted as additional evidence in support of the idea that the ranking statistics may be sensitive to the effect of impact over several years. Without

further research, however, this interpretation is not certain. the lack of significant correlation in these schools (and in 9-12 schools as well) may simply demonstrate the independence of the ranking statistics with absolute achievement level.

Discussion

The procedures discussed in this paper currently are being employed by the Dallas schools in a program to recognize and reward outstanding schools. Under this program, teachers in schools ranked in the top quarter of each grade-level category of schools will each receive a stipend of $1500. Other employees in those schools also will receive stipends, the amount determined by position and responsibility.

Both to encourage cooperation among district employees and to respond to queries and objections regarding the program's operation, a committee of teachers and other individuals representing various groups was established nine months ago to generally oversee the program. This oversight committee meets once a week to review methodology and procedures and to voice and respond to concerns of their constituents. After having reviewed the results of the simulations with 1983 data, this committee has deemed that the lower bound statistic based upon individually standardized gain scores (LB_{δ}) will be used this spring to rank schools in the Dallas district.

There are two additional variables that serve as qualifying criteria in establishing eligibility to receive stipend awards. These variables, the use of which if successful will subsidize the program, are student and teacher attendance. Before a school can qualify for an award, its

student percent average daily attendance (%ADA) must at least equal its previous year %ADA or equal the district-wide %ADA for comparable schools in the current year. In addition, the school's median teacher-days absent must either be less than five days or be less than 90 percent of its previous year's rate. Initially all schools will be ranked. Then the oversight committee will disqualify schools on the basis of the attendance criteria.

Besides providing an objective procedure for measuring school effectiveness, the over-all program is perceived as having a number of practical advantages. First it is designed to foster teamwork within schools and healthy competition among schools. Although many of its detractors view the program as a merit pay system it is really not such a system, at least not in the way these systems are generally conceived. The program is not designed to reward individual competition among teachers. In fact the authors would hesitate to recommend the use of similar procedures to evaluate individual teachers (the tests are not sensitive enough for this purpose). However, ten years of experience with test results suggest that they can be used to evaluate schools. Better schools do tend to achieve higher gains in student achievement.

A second advantage of the program is that it focuses on what district teachers, administrators, and popular constituency view as the most important outcome of instruction. To the extent the program is successful, achievement will increase in Dallas.

A third advantage is that the procedures described afford all schools an opportunity to be distinguished independently of where on the achievement continuum their student population began. The technique

rewards those schools that move the largest proportion of their students the farthest.

This paper has discussed a quantitative procedure for measuring school effectiveness. The program is norm-referenced in the sense that schools are compared with each other and ranked. Not all schools will win stipends for their teachers. It has been demonstrated that the procedures described serve their intended purpose.

References

Anderson, L.M. et. al. "Principles of Small Group Instruction in Elementary Reading." East Lansing, Michigan: Institute for Research on Teaching, College of Education, Michigan State University, Unpublished paper, 1982.

Block, James H. "The Effects of Various Levels of Performance on Selected Cognitive, Affective, and Time Variables." Ph.D. dissertation, University of Chicago, 1970.

Bloom, Benjamin. Human Characteristics and School Learning. New York: McGraw-Hill Book Company, 1976.

Brookover, W.B., & Lezotte, L.W. Changes in School Characteristics Coincident with Changes in Student Achievement. East Lansing, Michigan: Institute for Research on Teaching, College of Education, Michigan State University, 1979. (ED 181 005).

Doyle, W. "Accomplishing Writing Tasks in the Classroom." Paper presented at the meeting of the American Educational Research Association, Los Angeles, 1981.

Duffy, G.G., Teacher Effectiveness Research Implications for the Reading Profession. East Lansing, Michigan: The Institute for Research on Teaching, College of Education, Michigan State University, 1980.

Dunn, Rita and others. "Learning Style Researchers Define Differences Differently," Educational Leadership, February 1981, 372-375.

Edmonds, Ronald, "Effective Schools for the Urban Poor." Educational Leadership, 37 (October 1979): 15-27.

Emmer, E.T., and Anderson, L.M. "Effective Classroom Management at the Beginning of the School Year." Elementary School Journal, 1980, pp. 219-231.

Gage, N.L. "The Scientific Basis of the Art of Teaching." New York: Teachers College Press, Columbia University, 1978.

Glenn, B.C., What Works? An Examination of Effective Schools for Poor Black Children. Cambridge, Massachusetts: Center for Law and Education, Harvard University, 1981.

Levine, D.U. & Stark, J., Extended Summary and Conclusions: Institutional and Organizational Arrangements and Processes for Improving Academic Achievement at Inner City Elementary Schools. (mimeo) Kansas City, Missouri: University of Missouri-Kansas City, School of Education, Center for the Study of Metropolitan Problems in Education, August 1981.

Lysakowski, Richard S., and Walbert, Herbert J., "Instructional Effects of

Cues, Participation and Corrective Feedback: A Quantitative Synthesis." American Educational Research Journal, 19 (Winter 1982): 559-578.

Peterson, Penelope L. "Direct Instruction: Effective for What and for Whom? Educational Leadership 37 (October 1979): 46-48.

Rosenshine, B. "Content, Time and Direct Instruction." In A. Walbert and P. Peterson (Eds.), Research on Teaching; Concepts, Findings and Implications Berkely: McCutchan Publishing, 1979.

Rosenshine, Borak, and Furst, Norma. "The Use of Direct Observation to Study Teaching," in Second Handbook of Research on Teaching. Travers (ed.), Chicago: Rand McNally College Pub. Co., 1973.

Sheehan, Daniel S. and Marcus, Mary M. "The Effects of Desegregation on Student Attitudes and Achievement," IR 78-231-3-13, Dallas Independent School District, August, 1978.

Slavin, Robert. "Synthesis of Research on Cooperative Learning." Educational Leadership Vol. 38, No. 8, pp. 655-660.

Stallings, Jane, "Allocated Academic Learning Time Revisited, or Beyond Time on Task." Educational Researcher, Vol. 9 (December 1980): 11-16.

Weber, G. Inner-City Children Can Be Taught to Read: Four Successful Schools. Washington, D.C.: Council for Basic Education, 1971.